DOCUMENT RESUME

ED 370 968 TM 021 519

AUTHOR Tyson, LeaAnn; Silverman, Stephen
TITLE An Analysis of Statewide Teacher Appraisal Scores across Four Years.
PUB DATE Apr 94
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; *Elementary School Teachers; Elementary Secondary Education; Evaluation Methods; Longitudinal Studies; School Districts; *Scores; *Secondary School Teachers; State Programs; *Teacher Evaluation; Testing Programs
IDENTIFIERS Large Scale Programs; Texas; *Texas Teacher Appraisal System

ABSTRACT

Teacher evaluation is receiving increasing attention. Texas is the largest state to adopt a statewide appraisal instrument. The purpose of this study was to examine differences in Texas Teacher Appraisal System (TTAS) scores over a period of four years. The sum of scores of the first four individual domains and the overall summary performance score for teachers in a large school district in central Texas for four years (1988-1989, 1989-1990, 1990-1991, and 1991-1992) were examined. Specifically, scores between appraiser types (primary or secondary), levels (elementary or secondary), and appraisal periods (first or second) across years were investigated. Overall, results showed that scores increased significantly over the four year period, there was a significant appraiser type effect (although significant differences between scores awarded by primary appraisers and second appraisers were not evident in the fourth year), there was no significant difference between elementary teachers and secondary teachers, and there was no significant difference between appraisal periods. (Contains 38 references and 5 tables). (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 370 968

AN ANALYSIS OF STATEWIDE TEACHER APPRAISAL SCORES ACROSS FOUR YEARS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LEAANN TYSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LeaAnn Tyson
Western Washington University

Stephen Silverman
University of Illinois at Urbana-Champaign

Send correspondence to:

LeaAnn Tyson
Department of Physical Education, Health and Recreation
Carver Gymnasium
Western Washington University
Bellingham, WA 98225
206/650-3054

Presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April 4-8, 1994

TM021519

BEST COPY AVAILABLE

ABSTRACT

Teacher evaluation is receiving increasing attention. Texas is the largest state to adopt a statewide appraisal instrument. The purpose of this study was to examine differences in Texas Teacher Appraisal System (TTAS) scores over a period of four years. The sum of scores of the first four individual domains and the overall summary performance score for teachers in a large school district in central Texas for four years (1988-1989, 1989-1990, 1990-1991, and 1991-1992) were examined. Specifically, scores between appraiser types (primary or second), levels (elementary or secondary), and appraisal periods (first or second) across years were investigated. Overall, results showed that scores increased significantly over the four year period, there was a significant appraiser type effect (although significant differences between scores awarded by primary appraisers and second appraisers were not evident in the fourth year), there was no significant difference between elementary teachers and secondary teachers, and there was no significant difference between appraisal periods.

ANALYSIS OF STATEWIDE TEACHER APPRAISAL SCORES ACROSS FOUR YEARS

Evaluating the effectiveness of personnel is an important and difficult task. The evaluation of teachers' work has gained increased attention from the media, policy makers, and practitioners in the past ten to fifteen years (Ellett, 1990). It is often assumed that a key to improving education is to upgrade the quality of teachers. Increasing concerns about the credibility and accountability of teachers have led educational reforms to focus on teacher assessment (Chauvin, Loup & Ellett, 1991; Crisci, March & Peters, 1991; Darling-Hammond, 1990; Poster, Poster & Bennington, 1991).

Many suggest that the intent of teacher evaluation is to improve instruction. However, assessment data can be used to make administrative decisions, such as teacher dismissal, contract renewal, merit pay, and career ladder placements. Although there are many issues involved in teacher assessment, such as number of evaluations per teacher per year and what data should be collected on each teacher, most people agree that assessment should take place and the evaluators should be trained and skilled.

Evaluating the effectiveness of teachers may involve many aspects of assessment such as competency testing, student achievement measures, peer evaluation, self-assessment, and examination of teacher-produced documents. Most agree that evaluation should not be based on a single assessment and that teacher observation is critical in evaluation. Consequently, observation of teaching is by far the most common method of teacher evaluation; although, it is generally not the sole method of collecting information for assessment, and recent efforts are focusing on student learning, as well as specific teacher behaviors (Amos & Cheeseman, 1991; Ellett, 1990; Ellett & Garland, 1986; Wise, Darling-Hammond, McLaughlin & Bernstein, 1984).

Early efforts in classroom observation involved the use of high-inference measurements. In the period of the late 1950's through the 1970's, however, research in

the classroom contributed to the development of many low-inference observation systems (Brophy & Good, 1986; Medley & Mitzel, 1963; Rosenshine & Furst, 1973). Because of increased demands for teacher competency and the availability of research and instruments, legislatively-enacted development of large scale teacher assessment systems followed in states such as Arkansas, Connecticut, Florida, Georgia, Kentucky, Mississippi, Missouri, North Carolina, South Carolina, Tennessee, Texas and Virginia (Chauvin & Ellett, 1991; Ellett, 1990). As many as eighteen states have used, at one time or another, evaluation systems that involve "on the job" assessment for purposes of teacher certification, merit pay, career ladders, and professional development (Association of Teacher Educators, 1988; Chauvin et al., 1991; United States Department of Education, 1987). In addition, many states are considering efforts of these types (Chauvin et al., 1991).

To date, the most populous state to implement state-wide teacher appraisal using observation is Texas. The Texas Teacher Appraisal System (TTAS) was based on, for the most part, research which appeared in the 1970's and 1980's. This instrument was implemented state-wide in 1986 and is based primarily on classroom teaching performance (Barnes, 1987; Texas Education Agency, 1988). The TTAS contains 5 domains, broken down into 13 criteria, which are further divided into 65 behavioral indicators (Texas Education Agency, 1988). The first four domains (Instructional Strategies, Classroom Management and Organization, Presentation of Subject Matter, and Learning Environment) are based on classroom teaching performance, and the fifth domain is related to professional growth. All teachers, regardless of subject area or grade level taught, are assessed using this instrument. Separate observations are conducted by a primary appraiser (the teacher's immediate supervisor) and a second appraiser (possibly another district administrator).

The Texas Career Ladder, a now defunct merit pay system of three levels, was directly related to the TTAS during its time of existence (from 1986 to 1993 -- including the years of this study). Although one program could exist without the other, the TTAS and

the career ladder were consistently linked because the performance score on the TTAS was one criterion for advancement and maintenance at the three career ladder levels. Years of teaching experience and hours spent in professional growth activities served as the other two criteria for Career Ladder placement (Texas Education Agency, 1988). For teachers at Level 2 or Level 3 of the Career Ladder, a monetary bonus was awarded. The state funded a portion of the merit pay, and the individual school districts were responsible for the remainder. Although the Texas Career Ladder no longer exists, the TTAS is still used for evaluation purposes.

With more and more teachers advancing on the career ladder, some districts found it necessary to adopt stricter TTAS performance criteria. In some districts, higher scores were required for placement and/or maintenance on the various levels, or scores were rank ordered and only a portion of teachers that qualified under state and district criteria were funded. This increased competition for career ladder placements contributed to the interest in scoring issues, such as subgroup differences in scores on the TTAS.

Previous research has indicated subgroup differences. On-campus evaluators/principals tend to award higher scores than off-campus evaluators, or those who do not serve as the teacher's supervisor (Cronin & Capie, 1986; Ellett & Capie, 1985; Ellett, Teddlie & Naik, 1991; Kelley, 1985; Rose & Huynh, 1984; Wise et al., 1984). Earlier research on the TTAS suggested that primary appraisers award higher scores (Holmes, 1989; Performance Assessment Systems, Inc., 1987; Tyson, 1991). It has been suggested that the principal (or immediate supervisor) may experience role conflict in trying to serve as the instructional leader and an individual responsible for making administrative decisions, while trying to respond to the social context of the campus (Duckett, 1985; Stanley & Popham, 1988).

Elementary teachers tend to receive higher scores than secondary teachers on teacher evaluation instruments (Alabama Career Incentive Program Project Staff, 1987; Dickson & Wiersma, 1984; Florida Coalition for the Development of a Performance

Measurement System, 1984; Ligon & Ellis, 1986; Tadlock & Nesbit, 1984). Pilot studies on the TTAS, research on the initial TTAS instrument (prior to revision in 1987) and research on the current TTAS instrument also indicate that elementary teachers receive higher scores (Holmes, 1989; Performance Assessment Systems, Inc., 1986; Tyson & Silverman, in press). The research base for most instruments involve studies conducted at the elementary level (Holdzkom, 1991), and it has been suggested that different teaching behaviors are more and less effective for students at different grade levels (Brophy & Good, 1986).

Prior research on the TTAS indicates that scores increase from the first appraisal period to the second, and from one year to the next (Tyson & Silverman, in press). This is to be expected because the instrument and appraisal process are intended to improve instruction. However, continual increases in scores across years yields little discrimination in quality of teaching performance among teachers.

Little is known about the differences of scores between subgroups of teachers or scoring trends across time. Consistent and significant subgroup differences and score increases over time may be problematic, particularly if scores are used for administrative decisions such as career ladder placements. Given the limited information available on large scale teacher evaluation, an examination of the TTAS provides a basis from which to consider state mandated teacher evaluation for other states, as well. Examining large scale evaluation systems as they are implemented is an important issue. Although teacher assessment research may be conducted in "artificial" settings by controlling and equalizing the number of subjects per group, research in actual school district settings, where scores are used to make important administrative decisions, is critical in understanding how policies and practices impact teachers. It is difficult to analyze data of this type, and the only way to examine the practical aspects of how teacher evaluation scores impact teachers is by post hoc evaluation.

The purpose of this study was to examine differences in Texas Teacher Appraisal System (TTAS) scores over a period of four years. The sum of scores of the first four individual domains (Instructional Strategies, Management and Organization, Presentation of Subject Matter, and Learning Environment) and the overall summary performance score for teachers in a large school district in central Texas for four years (1988-1989, 1989-1990, 1990-1991, and 1991-1992) were examined. Specifically, scores between appraiser types (primary or second), levels (elementary or secondary), and appraisal periods (first or second) across years were investigated.

Method

Subjects

The subjects (N=620) were teachers in a large school district in central Texas for the school years 1988-1989, 1989-1990, 1990-1991, 1991-1992. Teachers in the study were included only if they taught in the district for each of the four years. All teachers were observed and evaluated with the TTAS by a certified appraiser, following the procedure mandated by the State Board of Education.

Texas Teacher Appraisal System

The instrument is categorized into five major domains which include the following:

1) Domain I -- Instructional Strategies
2) Domain II -- Classroom Management and Organization
3) Domain III -- Presentation of Subject Matter
4) Domain IV -- Learning Environment
5) Domain V -- Professional Growth and Responsibilities

Within each domain are broad criteria (a total of 13 criteria for the 5 domains) which are further broken down into behavioral indicators (a total of 65 indicators). When the instrument is scored, each indicator is worth one point and each criterion in Domains I through IV has three "exceptional quality" points that may be awarded at the criterion level.

A total of 76 points is available in Domains I through IV, and a total of 92 points is possible in the entire instrument.

For the years of this study, teachers on Levels 2 or 3 of the Texas Career Ladder were required to have two observations per year (typically one per semester), one by the primary appraiser (the teacher's supervisor) and one by the second appraiser (another campus administrator or off-campus administrator), as long as the teacher's performance was higher than satisfactory. Teachers on Level 1 of the career ladder were required to have four observations per year -- half of which were performed by the primary appraiser and half of which were performed by the second appraiser. One observation by each appraiser was conducted in each of the two appraisal periods, which paralleled the fall and spring semesters. Although teachers of equal or higher career ladder levels were allowed to serve as appraisers for other teachers (State Board of Education Rule sec. 149.43, 1990), the district in this study utilized only district administrators as appraisers.

Domains I through IV of the TTAS were scored by both appraisers and reflected teacher behavior during the observation period. Domain V was scored only by the primary appraiser and involves job-related behaviors that are not directly observable in the lesson. An example of such a behavior is "Progresses in growth requirements or none needed."

In the school district studied, the appraisal process differed at the elementary level and the secondary level. At the elementary level, the primary appraiser was the campus principal, and the second appraiser was an off-campus administrator or district appraiser (the district employed two full-time appraisers). At the secondary level, the primary supervisor was the campus principal or assistant principal (grade level principal), and the second appraiser was generally another on-campus administrator (except in certain specialized areas, such as fine arts or special education).

Appraisers scheduled and observed 45-minute lessons independently. Scoring was done on each observation, yet all scores of all observations were used to calculate the overall summary performance score for each teacher. This summative score was weighted,

with 60% and 40% of the overall summary performance score accounted for by the primary appraiser and second appraiser, respectively.

Prior to certification as an appraiser, individuals completed approved training, passed a test on procedures, passed a scoring proficiency test on a videotaped lesson, and performed a field test with at least two other appraisers on an actual lesson. Yearly updates and recertification every three years are required for appraisers to maintain their certifications. All appraisers in this study were certified and all procedures followed those mandated by the State Board of Education.

Appraisal Process

Teachers received TTAS orientations prior to observations being conducted. Teachers also received yearly updates on the procedures and TTAS instrument. Prior to each observation conducted during the school year, teachers participated in a pre-observation conference with the appraiser. Following the observation, the teacher and appraiser met for a post-conference to review the teacher's score or performance on that observation. Teachers could appeal the appraiser's procedure through the local grievance procedure. Information regarding the teacher's overall summary performance score and career ladder status was presented by the primary appraiser at the summative conference at the conclusion of the appraisal year. Primary appraisers were responsible for inputting all appraisal data into the district computer files.

Integrating Data

Data from the school district were contained on the Appraisal Record File and the Monthly Employee File for each of the years under study. The Appraisal Record File for each year contained the following information: teacher social security number; appraiser social security number; whether the appraiser was the primary or second appraiser; which observation (first or second) by the appraiser; individual domain scores; and overall summary performance score. The information in the Appraisal Record File was input by primary appraisers after all observations of each teacher were completed in any given year.

In addition, the school district maintained separate files with demographic information for each teacher. The Monthly Employee Files contained social security number, campus number, job code, and career ladder level.

The school district granted permission to allow the employee information and appraisal data to be used for research purposes. Files were accumulated for the school years 1988-1989, 1989-1990, 1990-1991, 1991-1992. Each file was transferred to another computer for subsequent analysis. During the transfer process, teachers' names were deleted from the files; therefore, teachers could be identified only by number.

Data Analysis

Only teachers who taught in the district each of the four years were included in the study. In instances where a subject was missing any necessary information in any of the four years, or the campus assignment was other than one of the the 26 campuses in the district, that subject subsequently was deleted from the analysis. In addition, if teachers changed subgroups (as when an elementary teacher moves to the secondary level) during the four years of study, that teacher's scores were included in the appropriate subgroup for any given year.

When examining the sum of Domains I through IV, separate 2X4 (level: elementary/secondary X year: 88-89/89-90/90-91/91-92) ANOVAs with repeated measures on the second factor were computed for teachers with two and four observations. Also for this same dependent variable, a 2X4 (appraiser type: primary/second X year: 88-89/89-90/90-91/91-92) ANOVA with repeated measures on both factors was computed for teachers with two observations, and a 2X4X2 (appraiser type: primary/second X year: 88-89/89-90/90-91/91-92 X period: first/second) ANOVA with repeated measures on all three factors was computed for teachers with four observations. To analyze the overall summary performance scores for all teachers, a 2X4 (level: elementary/secondary X year: 88-89/89-90/90-91/91-92) ANOVA with repeated measures on the second factor was computed. For all analyses, alpha was set at .01.

Results

Descriptive Information

For 1988-1989, all teachers had a mean overall summary performance score of 168.41. For the second year, the mean was 172.13. For the last two years (1990-91 and 1991-92), the mean overall summary performance scores were 171.29 and 174.21, respectively.

Differences Between Appraiser Types (Primary or Second) Across Years

Sum of the first four domains -- Instructional Strategies, Management and Organization, Presentation of Subject Matter, Learning Environment. For teachers with two observations, there was a significant difference between years ($F[3, 1614] = 61.59$, $p<.01$), with scores tending to be higher each year, for the most part. There also was a significant difference between appraiser type ($F[1,538] = 54.06$, $p<.01$), with higher scores awarded by the primary appraisers for each year. In addition, there was a significant interaction (appraiser type X year) effect ($F[3, 1614] = 6.91$, $p<.01$). Means of the sum of Domains I through IV by appraiser type and years for teachers with two observations are presented in Table 1.

Insert Table 1 about here

Follow-up Student-Newman-Keuls tests identified many significant differences between the means of the sum of Domains I through IV. For each year, the scores awarded by the primary appraisers were significantly higher, except in year 4. When examining increases in scores, primary appraisers' scores increased significantly from year 1 to year 2, and from year 3 to year 4. Scores awarded by the primary appraiser dropped slightly (not significantly) from year 2 to year 3. Second appraisers' scores followed the same pattern, except that the decrease in scores from year 2 to year 3 was significant. In addition, many of the various means were significantly different from each other. For

example, the primary appraisers' scores in year 4 were significantly higher than the second appraisers' scores in year 1. Because the focus of this analysis was the difference in appraiser type by year, these differences were not meaningful in this study.

For teachers with four observations, there was a significant difference in overall means between years, with scores increasing each of the four years ($F[3, 240] = 49.95$, $p<.01$). Additionally, there was a significant difference between appraiser type ($F[1,80] = 52.77$, $p<.01$), with the higher scores being awarded by the primary appraisers in each year. There was no significant interaction effect. Means of the sum of Domains I through IV by appraiser type and years for teachers with four observations are presented in Table 2.

Insert Table 2 about here

Follow-up analysis indicated significant increases in the overall means from year to year, except from year 2 to year 3. When examining increases by appraiser type, the scores awarded by the primary appraisers increased significantly from year to year, except from year 2 to year 3. Second appraisers scores increased significantly from year 1 to year 2 and from year 3 to year 4, but decreased slightly (not significantly) from year 2 to year 3.

Differences Between Levels (Elementary or Secondary) Across Years

Sum of the first four domains -- Instructional Strategies, Management and Organization, Presentation of Subject Matter, Learning Environment. For teachers with two observations, there was no significant difference between the scores of elementary and secondary teachers. There was a significant years effect, however ($F[3, 3228] = 69.39$, $p<.01$). In addition, there was a significant interaction effect ($F[3, 3328] = 17.93$, $p<.01$). Mean scores of teachers with two observations by level and year are presented in Table 3.

Insert Table 3 about here

Student-Newman-Keuls follow-up tests identified many significant differences among the means. For elementary teachers, scores increased significantly from year 1 to year 2, and from year 3 to year 4. Scores decreased significantly from year 2 to year 3. Secondary teachers' scores increased each year, with the differences significant from year 1 to year 2 and from year 3 to year 4. In years 1 and 2, the scores of the elementary teachers were significantly higher. In the final two years, secondary teachers' scores were slightly higher, although not significantly.

For teachers with four observations, there was no significant difference between scores of elementary and secondary teachers. However, in each year, elementary teachers received higher scores. There was a significant years effect ($F[3, 966] = 94.03$, $p<.01$), with scores increasing each year. Follow-up tests showed that scores increased significantly from year to year, except from year 2 to year 3. There was no interaction effect. Scores by level and year for teachers with four observations appear in Table 4.

Insert Table 4 about here

Overall summary performance score. When teachers' overall summary performance scores (the summative score which has been weighted in the primary appraiser's favor) were examined, no significant difference between elementary and secondary teachers' scores was seen. There was, however, a significant years effect ($F[3, 1854] = 94.77$, $p<.01$) and a significant interaction effect ($F[3, 1854] = 16.46$, $p<.01$). Means of the overall summary performance scores of all teachers by level and year are presented in Table 5.

Insert Table 5 about here

Follow-up tests indicated that elementary teachers' scores increased significantly from year 1 to year 2 and from year 3 to year 4, but decreased significantly from year 2 to year 3. Scores for secondary teachers increased each year, but the increases were significant from year 1 to year 2 and from year 3 to year 4 only. In year 1 and year 2, the differences between elementary and secondary teachers were significant; however, no significant differences were seen between levels in year 3 and year 4.

Differences Between Appraisal Periods (First or Second) Across Years

Sum of the first four domains -- Instructional Strategies, Management and Organization, Presentation of Subject Matter, Learning Environment. For this analysis, only teachers with four observations were included. Teachers with two observations receive one observation in each appraisal period by different appraisers (primary and second). Those teachers with four observations receive one observation by the primary appraiser and one observation by the second appraiser in each of the two appraisal periods. Therefore, the two appraisal periods may be compared. In this analysis, there was no significant difference in appraisal period and there was no interaction of appraisal period and year, or appraiser type and appraisal period.

Discussion

It should be noted that the scores were extremely high. Although there were many significant differences found, some of these differences involved less than one point. Each indicator in the instrument is worth one point, and each exceptional quality award is worth three points. Thus, the significant differences found often were less that the difference between receiving or not receiving a standard expectation point on an indicator or exceptional quality credit on one criterion. When teachers receive the summative score, they are labeled in categories ranging from "Unsatisfactory" to "Clearly Outstanding." The overall summary performance score required to receive a "Clearly Outstanding" rating is 160. As can been seen, the overall summary performance score means were all higher. It is highly unlikely that all or most teachers were, indeed, "clearly outstanding," although the

district (like most) prided themselves in having very strong teachers. Despite the high scores, significant differences in subgroups become important when teachers' scores are rank ordered and only a portion of teachers are funded with Career Ladder bonuses.

It is also interesting to note that the school district under study had additional appraiser training and inservice between the second and third school year. At that time, appraisers were presented with information on scores and scoring trends of the previous two years. In several of the analyses in the current study, significant increases in scores were not seen between year 2 and year 3. In a few instances, scores decreased.

Differences Between Appraiser Types (Primary or Second) Across Years

For the most part, scores increased over the four year period and primary appraisers awarded higher scores. The increase is to be expected if the appraisal process is meeting its primary purpose -- to improve instruction. Higher on-campus or primary appraiser scores are consistent with the literature on large scale evaluation and the TTAS (Cronin & Capie, 1986; Ellett & Capie, 1985; Ellett et al., 1991; Holmes, 1989; Kelley, 1985; Rose & Huynh, 1984; Tyson, 1991; Wise et al., 1984). It is often thought that the an on-campus evaluator is more familiar with the teacher's overall abilities and may actually score the teacher on overall performance, rather than on just the formal observation (Holmes, 1989). Second appraisers, on the other hand, have no basis for awarding scores other than the formal observation because they do not supervise the teacher directly. The primary appraiser's greater familiarity could suggest that the score may be more dependable (Doyle, 1983), or it may suggest that the second appraiser's score may be more objective (Association of Teacher Educators, 1988).

The primary appraiser is in a position where he or she interacts with the teacher frequently, often on a daily basis because both the teacher and appraiser are on the same campus. The second appraiser visits the campus infrequently. The primary appraiser's frequent interaction and "face to face" contact could impact how the teacher is scored. After implementation of the TTAS, principals expressed concern about how the instrument,

particularly the greater formality, hampered relationships with teachers (Bezdek, Cross & Guerrero, 1988). Second appraisers may not be concerned with how the evaluation affects their relationship with the teacher since they do not deal with the teacher on a day to day basis. Role conflict of the principal, who must be the "helper person" for instructional improvement and the "hatchet person" when making administrative decisions, may also contribute to the difference in scores of the two appraiser types (Ducket, 1985; Stanley & Popham, 1988).

It might be assumed that superior teachers will receive similar scores from different appraisers on different occasions, and likewise, weaker teachers will be scored similarly by both appraisers. In studies where comparisons are made when teachers were observed on the same day and when teachers were observed on separate occasions, variance due to occasions was greater than variance due to observers who saw the same lesson (Cronin & Capie, 1986). Variation due to teachers when observed on different days has been as much as three times that of teachers observed on the same day (Yap & Capie, 1985). Thus, variation in performance can be expected. However, when teachers receive scores from one appraiser that are higher than those from another, the appraiser who awarded the lower score is often assumed to be "inaccurate," and it has little to do with instability of performance. Significant differences between scores awarded by primary and second appraisers are problematic for teachers.

As can be seen, scores increased significantly during the four year period. Scores which are already relatively high, and continue to increase can be problematic for any school district. It would be hoped that scores would increase as instructional improvement occurs; however, when there is competition for limited funding of bonuses, appraisers may find themselves feeling obligated or pressured into giving higher scores. Instructional improvement may be interpreted in several ways. Teachers may be performing better for the appraisers, particularly after the appraiser has conferenced and perhaps suggested ways to improve. Teachers may become more familiar with the instrument and what is expected

for each indicator and criterion over the years. Appraisers may feel obligated to award higher scores each year because teachers can and do improve. As noted, scores did not increase significantly from year 2 to year 3, when additional appraiser staff development occurred.

There were no significant differences between appraiser types in the fourth year for teachers with two observations. This, along with the significant interaction effect, suggests that scores between appraiser types are "getting closer." The differences between primary appraiser and second appraiser scores decrease over time. Because scores are also increasing, the decreasing differences suggest that there is a ceiling effect of already high scores. Primary appraiser scores cannot increase as much as second appraisers scores across four years because primary appraiser scores began very high.

Differences Between Levels (Elementary or Secondary) Across Years

When examining teachers with two and four observations for both dependent variables (sum of Domains I through IV and the overall summary performance score), there were no significant differences between elementary and secondary teachers. This does not support other literature on large scale evaluation, where elementary teachers tend to receive higher scores (Alabama Career Incentive Program Project Staff, 1987; Dickson & Wiersma, 1984; Florida Coalition for the Development of a Performance Measurement System, 1984; Holmes, 1989; Ligon & Ellis, 1986; Performance Assessment Systems, Inc., 1986; Tadlock & Nesbit, 1984; Tyson & Silverman, in press). However, there has been little, if any, research examining the differences over time. The significant years effect again demonstrates the increasing scores. Scores of elementary and secondary teachers increased significantly from year 1 to year 2 and from year 3 to year 4. As mentioned previously, the lack of a significant increase or the decrease in scores from year 2 to year 3 may have been due to the appraiser training that occurred at that time.

The significant interaction effect, when examining the sum of the first four domains for teachers with two observations and the overall summary performance score for

all teachers, shows secondary teachers' scores are increasing at a greater rate. Because elementary teachers tended to have higher scores, this suggests that the differences between these two subgroups are becoming less over time. Here again, the ceiling effect of already high scores could be occurring. Although scores of all teachers were initially high, secondary teachers' scores have "more room" to improve.

Implications and Importance of the Study

Limited research exists on large scale teacher evaluation. In addition, virtually no research has been conducted examining evaluation scores over time. Prior research on the TTAS showed similar results to studies on other instruments, with elementary teachers receiving higher scores and primary appraisers, or principals, awarding higher scores. In this study, similar trends emerged. However, subgroup differences were not significant as scores were analyzed over a period of four years. Because scores continue to increase from year to year, it is logical that subgroup differences would no longer occur as a ceiling effect is seen. Thus, if consistency in scores among subgroups may seem desirable, the consistency could be attributed to scores that have reached a ceiling.

The results of this study have implications in teacher evaluation. Because a primary purpose of teacher evaluation is instructional improvement, it would be expected that teachers would improve over time. Teacher assessment scores that continue to rise from year to year may be indicative of instructional improvement. If scores are used for administrative decisions, however, little distinction among teachers will exist and the utility of the instrument may be weakened, particularly if scores are already relatively high. As previously noted, scores from year 2 to year 3 did not increase as much or decreased when compared to the other years. The additional inservice and training of appraisers may have contributed to this. If high scores are viewed as problematic, it appears that staff development is beneficial.

Although subgroup differences tend to be problematic because they suggest that lower scoring teachers are in need of staff development, appraisers are in need of additional

training, or validity across grade levels should be examined, consistency of scores among subgroups also is problematic if all subgroups are approaching the highest possible score. Those implementing large scale evaluation should take a close look at how subgroup differences are interpreted and handled.

Because educational reform efforts tend to focus on the quality of teachers, teacher evaluation will continue to be important. Research to determine if teachers improve through assessment should be conducted, and statewide appraisal scores should continue to be examined. The real benefit of research will be when it is possible to determine if teachers with higher evaluation scores produce the greatest student learning. Additional research on teacher evaluation and scoring trends across years can only benefit-policy makers, those implementing teacher evaluation systems, and those considering large scale assessment systems.

References

Alabama Career Incentive Program Project Staff. (1987). Alabama performance-based career incentive program: A report on the norming study of the spring semester. Atlanta, GA: Georgia State University. (ERIC Document Reproduction Service No. ED 302 578)

Amos, N., & Cheeseman, R. H. (1991, November). An analysis of provisional teachers failing the Mississippi assessment instruments for certification. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Lexington, KY. (ERIC Document Reproduction Service No. ED 341 669).

Association of Teacher Educators. (1988). Teacher assessment. Reston, VA: Author. (ERIC Document Reproduction Service No. ED 289 869)

Barnes, S. (1987, April). The development of the Texas Teacher Appraisal System. Paper presented a the Annual Meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 294 323)

Bezdek, R., Cross, R., & Guerrero, T. (1988, April). Bureaucratic control and principal role. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 303 897)

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed.). New York: Macmillan.

Chauvin, S. W., & Ellett, C. D. (1991, April). Replacing lifetime certification with a renewable credential: A survey of Louisiana educators' perceptions of the Louisiana teaching internship and statewide teacher evaluation programs. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335 352)

Chauvin, S. W., Loup, K. S., & Ellett, C. D. (1991, April). Development and validation of a comprehensive assessment system for teaching and learning. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago. (ERIC Document Reproduction Service No. ED 335 410)

Crisci, P. E., March, J. K., & Peters, K. H. (1991, April). Using the current paradigms in teacher training to prepare principals and mentor teachers to appraise classroom instruction. Paper presented at the Annual Meeting of the American Association of Colleges of Teacher Education, Atlanta. (ERIC Document Reproduction Service No. ED 333 544)

Cronin, L. L., & Capie, W. (1986, April). The influence of daily variation in teacher performance on the reliability and validity of assessment data. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 274 704)

Darling-Hammond, L. (1990). Teacher evaluation in transition: Emerging roles and evolving methods. In J. Millman & L. Darling-Hammond (Eds.), Teacher evaluation. Assessing elementary and secondary school teachers (pp. 17-32). Newbury Park, CA: Corwin Press.

Dickson, G. E., & Wiersma, W. (1984). Empirical measurement of teacher performance. Research and evaluation in teacher education. Toledo, OH: Toledo University, Center for Educational Research and Services. (ERIC Document Reproduction Service No. ED 263 211)

Doyle, K. (1983). Evaluating teaching. Toronto: D. C. Heath.

Duckett, W. R. (Ed.). (1985). The competent evaluator of teaching: A CEDR monograph. (Report No. ISBN-0-87367-720-X). Bloomington, IN: Phi Delta Kappa, Center on Evaluation and Research. (ERIC Document Reproduction Service No. ED 266 536)

Ellett, C. (1990). A new generation of classroom-based assessments of teaching and learning: Concepts, issues and controversies from pilots of the Louisiana STAR. Baton Rouge, LA: Teaching Internship and Statewide Teacher Evaluation Projects, College of Education, Louisiana State University.

Ellett, C. D., & Capie, W. (1985, April). Assessing meritorious teacher performance: A differential validity study. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 264 255)

Ellett, C. D., & Garland, J. S. (1986). Teacher evaluation practices in our largest school districts: Are they measuring up to "state-of-the-art" systems? Baton Rouge, LA: Author. (ERIC Document Reproduction Service No. ED 294 316)

Ellett, C. D., Teddlie, C., & Naik, N. (1991, April). The effects of high stakes certification demands on the generalizability and dependability of a classroom-based teacher assessment system. Presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335 408)

Florida Coalition for the Development of a Performance Measurement System. (1984). Teacher evaluation study. Norming study, generic competence revision, feedback proposal, cognitive examinations, and specialized domains. Report for 1983-1984. Tallahassee, FL: Author. (ERIC Document Reproduction Service No. ED 266 122)

Holdzkom, D. (1991). Teacher performance appraisal in North Carolina: Preferences and practices. Phi Delta Kappan, 72, 782-785.

Holmes, G. E. (1989). The effects of appraiser, contextual, and procedural variables on the performance rating of teachers as determined by the Texas Teacher Appraisal System . Dissertation Abstracts International, 49, 2470A-2471A. (University Microfilms No. 8819501)

Kelley, M. F. (1985, April). The Arizona performance-based teacher certification program. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 263 078)

Ligon, G., & Ellis, J. (1986, April). Adjusting for rater bias in teacher evaluations: Political and technical realities. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 273 677)

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching (pp. 247-328). Chicago: Rand McNally.

Performance Assessment Systems, Inc. (1986). A summary analysis of the Texas Teacher Appraisal System (TTAS): Observations in six pilot districts during 1985. Athens, GA: Author.

Performance Assessment Systems, Inc. (1987). An interpretive summary of 1986-1987 TTAS data. Athens, GA: Author.

Poster, C., Poster, D., & Benington, M. (1991). Teacher appraisal. A guide to training. London: Routledge.

Rose, J. S., & Huynh, H. (1984, April). Technical issues for adopting the APT for districtwide teacher evaluation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans. (ERIC Document Reproduction Service No. ED 247 266)

Rosenshine, R., & Furst, N. (1973). The use of direct observation to study teaching. In M. W. Travers (Ed.), Second handbook of research on teaching (pp. 122-183). Chicago: Rand McNally.

Stanley, J. J., & Popham, J. W. (Eds.). (1988). Teacher evaluation: Six prescriptions for success. Alexandria, VA: Association for Supervision and Curriculum Development. (ERIC Document Reproduction Service No. ED 299 683)

Tadlock, J., & Nesbit, L. (1984, November). The relationship of teacher evaluation scores generated by a process-product evaluation instrument to selected variables. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. 252 931)

Texas Education Agency. (1988). Teacher orientation manual 1988-1989. Austin, TX: Author.

Texas State Board of Education Rule, sec. 149.41-149.44 (1990).

Tyson, L. A. (1991). Differences in Texas Teacher Appraisal System scores across appraisers, campuses, years, and various subgroups of teachers (Doctoral dissertation, The University of Texas, 1991). Dissertation Abstracts International, 52, 2508A.

Tyson, L. A., & Silverman, S. (in press). An analysis of physical education and non-physical education teachers at the elementary and secondary level on statewide teacher assessment. Journal of Teaching in Physical Education.

United States Department of Education. (1987). What's happening in teacher testing. An analysis of state teacher testing practices. Washington, DC: Author.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1984). Teacher evaluation: A study of effective practices. Santa Monica, CA: Rand Corporation. (ERIC Document Reproduction Service No. ED 246 559)

Yap, K. C., & Capie, W. (1985, March). The influence of same day or separate day observations on the reliability of assessment data. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 265 166)

Table 1.

Means of the Sum of Domains I Through IV
by Appraiser Type and Year

Teachers With Two Observations

(n = 539)

Year	Appraiser Type	Mean	Year Mean
1 (1988-89)	Primary	69.78	68.99
	Second	68.20	
2 (1989-90)	Primary	71.05	70.60
	Second	70.15	
3 (1990-91)	Primary	70.81	69.98
	Second	69.14	
4 (1991-92)	Primary	71.59	71.34
	Second	71.20	
	Points Available	76.00	

*Significant appraiser type effect, significant years effect, and significant interaction effect ($p<.01$).

*In each year, there was a significant difference ($p<.01$) between the primary and second appraisers' scores, except in year 4.

Table 2.

Means of the Sum of Domains I Through IV
by Appraiser Type and Year

Teachers With Four Observations

Year	Appraiser Type	Mean (n = 81)	Appraiser Type Mean (n = 162)	Year Mean (n = 324)
1 (1988-89)	Primary (1st)	61.15	62.51	61.95
	Primary (2nd)	63.88		
	Second (1st)	60.09	61.39	
	Second (2nd)	62.69		
2 (1989-90)	Primary (1st)	64.88	66.15	66.04
	Primary (2nd)	67.43		
	Second (1st)	64.84	65.92	
	Second (2nd)	67.00		
3 (1990-91)	Primary (1st)	66.51	66.88	66.20
	Primary (2nd)	67.26		
	Second (1st)	64.81	65.51	
	Second (2nd)	66.21		
4 (1991-92)	Primary (1st)	68.27	69.06	68.83
	Primary (2nd)	69.84		
	Second (1st)	67.98	68.61	
	Second (2nd)	69.25		
	Points Available	76.00		

*Significant appraiser type effect and significant years effect ($p<.01$).

Table 3.

Means of the Sum of Domains I Through IV
by Level and Year

Teachers With Two Observations

(Elementary n = 540, Secondary n = 538)

Year	Level	Mean	Year Mean
1 (1988-89)	Elementary	69.63	68.99
	Secondary	68.36	
2 (1989-90)	Elementary	71.14	70.60
	Secondary	70.06	
3 (1990-91)	Elementary	69.59	69.98
	Secondary	70.35	
4 (1991-92)	Elementary	71.19	71.34
	Secondary	71.61	
	Points Available	76.00	

*Significant years effect and significant interaction effect ($p<.01$).

Table 4.

Means of the Sum of Domains I Through IV
by Level and Year

Teachers With Four Observations

(Elementary n = 152, Secondary n = 172)

Year	Level	Mean	Year Mean
1 (1988-89)	Elementary	62.83	61.95
	Secondary	61.17	
2 (1989-90)	Elementary	66.81	66.04
	Secondary	65.35	
3 (1990-91)	Elementary	66.47	66.20
	Secondary	65.95	
4 (1991-92)	Elementary	69.49	68.83
	Secondary	68.25	
	Points Available	76.00	

*Significant years effect ($p<.01$).

Table 5.

Means of the Overall Summary Performance Score
by Level and Year

All Teachers

(Elementary n = 308, Secondary n = 312)

Year	Level	Mean	Year Mean
1 (1988-89)	Elementary	170.01	168.41
	Secondary	166.83	
2 (1989-90)	Elementary	173.49	172.13
	Secondary	170.78	
3 (1990-91)	Elementary	171.64	171.29
	Secondary	170.94	
4 (1991-92)	Elementary	174.07	174.21
	Secondary	174.34	
	Points Available	184.00	

*Significant years effect and significant interaction effect ($p<.01$).